Centre of Excellence for Development Impact and Learning

AUGUST 2022

CEDIL Methods Working Paper 9

Evaluation and Measurement

Orazio P. Attanasio
Elisa Cavatorta

LIDC London International Development Centre
UKaid from the British people
Oxford Policy Management

About CEDIL

The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by the UK government through UK Aid. The mission of the Centre is to test innovative methodologies in evaluation and evidence synthesis and promote evidence-informed development. CEDIL-supported projects fall into three programmes of work: evaluating complex interventions, enhancing evidence transferability, and increasing evidence use.

CEDIL methods working papers

The CEDIL methods working paper series offers innovative research methods to develop impact evaluation and evidence synthesis work in low- and middle-income countries.

About this working paper

The content of this paper is the sole responsibility of the authors and does not represent the opinions of CEDIL or the Foreign, Commonwealth & Development Office. Any errors and omissions are also the sole responsibility of the authors. Please direct any comments or queries to the corresponding authors, Orazio Attanasio at orazio.attanasio@yale.edu and Elisa Cavatorta at elisa.cavatorta@kcl.ac.uk.

Suggested citation: Attanasio, O. and Cavatorta, E. 2022. *Evaluation and Measurement*. CEDIL Methods Working Paper 9. London and Oxford: Centre of Excellence for Development Impact and Learning (CEDIL). https://doi.org/10.51744/CMWP9

Cover design: PhilDoesDesign

CEDIL methods working paper: Evaluation and Measurement

Authors

Orazio P. Attanasio
Yale University, Institute for Fiscal Studies, FAIR - Centre for Experimental Research on Fairness, Inequality and Rationality @NHH, National Bureau of Economic Research, and Center for Economic and Policy Research

Elisa Cavatorta
King's College London

Centre of Excellence for Development Impact and Learning

UKaid from the British people

Contents

Abstract iv

Executive summary v

1. Introduction 1

2. From decision-making to measurement: a conceptual framework 3

3. What to measure? 7

   3.1 Measuring outcomes 7

   3.2 Measuring to understand mechanisms 8

4. How to use measures and how to measure impacts 12

5. How to construct new measures 25

6. Conclusions 27

References 29

Abstract

Measurement is crucial to evaluation. Without appropriate measures, we cannot assess the effectiveness of policy. In this latest CEDIL Methods Working Paper, the authors address several conceptual issues linked to the construction and validation of appropriate measures that are useful for evaluation. In particular, the paper discusses the following issues: what to measure, how to measure it, how to use existing measures and how to construct new measures.

Executive summary

In evaluation work it is often the case that most attention is given to the direct outcomes of an intervention, using measures taken from other contexts or previous studies. However, if a goal of policymakers or researchers is improving the intervention when necessary, scaling it up, or answering questions about why and for whom it works, a focus solely on outcomes should be resisted. One challenge to this endeavour is that many of the drivers of behavioural changes that determine the success or failure of interventions are not directly or immediately observable. Sometimes measures do not exist, or it is not obvious that measures developed in other contexts are applicable.

This paper's goal is to offer a number of critical reflections on four key core questions for any evaluator: what to measure, how to measure it, how to use existing measures, and how to construct new measures. We argue that thinking about a conceptual framework that links human decision-making processes and measurement can provide a specific and useful angle that stresses the importance of identifying – using different techniques and measurement tools – the causal links that are relevant for a coherent use, and possible improvements, of the theories of change that are used today. We discuss the challenges and important considerations related to this process and the process of creating new or better measures. Measurement innovation will be useful in many disciplines and is an important area of research that should be promoted and supported.

This methodological paper has been written by two economists who have a cross-discipline interest in evaluation theory and methods, drivers of behaviour change, and measurement innovation.

1. Introduction

When evaluating a policy intervention, which will typically be designed to achieve a certain policy objective, the temptation is to focus the measures deployed in the field on the specific outcomes of interest in each evaluation exercise. It is often considered that the outcome of interest can be easily measured so as to establish the impact of the interventions being studied. However, such an approach is very narrow and reductive, and this temptation should be resisted.

Much debate has recently been devoted to different approaches to evaluation. In this paper, we argue that evaluation, regardless of the approach used in its execution, relies in a fundamental way on measurement, and that measurement should be approached very broadly and not limited to the outcomes of interest. The main reason for adopting such an approach is that, when evaluating a policy, it is crucial to understand the mechanisms that generate the observed impacts. In many disciplines this is evident by the emphasis now placed on the need for a 'theory of change'. Economists, who work on modelling individual behaviour and how that behaviour reacts to specific incentives and factors, can provide useful models that can be used to identify empirically the causal links that form a 'theory of change'. However, the identification of such links requires data that are sufficiently detailed and rich.

There are several reasons not to confine measurement solely to the outcome of interest. First, in most cases it is not immediately obvious what outcomes are affected by an intervention and, more generally, the variables that are informative about the policymakers' interest and that can improve our understanding of the theory of change for a specific intervention. Second, it is not always obvious what the appropriate metric is for evaluating the impact of an intervention; is a certain observed impact small or large? Third, many interventions are likely to affect multiple outcomes and policymakers should be interested in the outcomes that an intervention might target explicitly, potential complementarities between these outcomes, and other outcomes that can constitute side-effects (sometime undesirable side-effects). Fourth, even when certain positive effects are measured in an evaluation, many factors will influence an intervention's impacts when that intervention is deployed at scale. In that respect, evaluations can provide useful evidence on coverage, fidelity, and costs, and on the drivers of these important components of an intervention's success (or failure).

More generally, to be useful for policy design, an evaluation needs to understand the mechanisms that yield the observed effects, in terms of how an intervention affects the behaviour of its recipients, how it affects the surrounding environment, and what challenges could arise when implementing it at scale. Only then an evaluation can be used to improve an intervention, or extrapolate the results to different contexts.

Consider, for example, the evaluation of the impacts of a primary school nutrition intervention which provides breakfasts or lunches to poor children. It is possible that the parents of these children will reduce the food they provide to the target children at home, perhaps so as to provide a higher nutrition intake to their siblings. Or consider the provision of additional (less qualified) teaching assistants to certain schools or nurseries. It is possible that these additional resources will be used effectively, therefore resulting in sizeable positive impacts on child development, or it is possible that they will trigger a re-allocation of teachers' time that could reduce these impacts and even cancel them. A mere focus on the direct outcomes of the intervention would miss the wider implications of the intervention. Therefore, it is important to model and understand the behaviour of the individuals (even those who are not the direct beneficiaries and who are involved only indirectly) involved in the intervention, and the drivers of their behaviours. Such an endeavour constitutes what, in many disciplines, is identified as a theory of change. While the need for a theory of change to interpret the results of an evaluation is quite widely accepted in many disciplines, a big challenge facing researchers and policymakers is the empirical identification and quantification of the causal links that define the model of individual behaviour – or, in other words, the structural effect that (potential) mediators can have on the outcomes of interest. Economists have paid particular attention to these identification issues. We argue that it is increasingly the case that measurement (and innovative measurement in particular) can be key to such an endeavour.

This paper's goal is to offer a number of critical reflections on four key core questions for any evaluator: (i) what to measure when executing evaluations; (ii) how to measure it; (iii) how to use existing measures; and (iv) how to construct new measures.

Naturally, measurement tools need to be adjusted to the specific needs of each investigation as these needs should drive what is measured and how to use such measures. Despite the different methodological evaluation traditions present in different disciplines, measurement is key to all evaluation exercises. And different approaches can be usefully used to gain new insights and measures that could be useful in different contexts. As economists by training, we are inevitably influenced by the quantitative tradition that considers modelling individual behaviour and reactions to individual incentives as central to establishing causal links. The identification of these causal links from the drivers of behaviour to observed choices and outcomes is difficult, as some of them are inherently unobservable. In what follows we stress the importance, for this reason, of engaging with and measuring constructs and variables that economists have been reluctant to use in the past and that have, instead, been used by other disciplines and fields (e.g. the measurement of perceptions, or beliefs about the future). We hope these reflections will be useful for a wide range of social scientists.

2. From decision-making to measurement: a conceptual framework

The conceptual framework that informs the design of policy interventions and their evaluation should include the outcomes of interest, the individual behaviours that determine them and their drivers of change, and the measurement of all of these factors. As we discuss below, many of these factors are often not directly observable, and yet they are of major interest to policymakers and researchers. What is needed is a formal model that links such constructs to the available measures and/or that informs the construction and design of specific measurement tools. In this sense, theory and measurement are inherently related, in that the latent variables of interest to the policymaker and their determinants should inform what measures are collected and how. And yet in most contexts they do not necessarily develop jointly.

The conceptual frameworks that are implicitly or explicitly used in social sciences when engaging in evaluation represent the formalisation of decision-making behaviour by individuals who are motivated by a set of objectives, social norms, attitudes, and beliefs. One of the main objectives of serious and deep evaluation in social sciences contexts should be understanding individual behaviour and how it is affected and influenced by a policy intervention. There are strong motivations for such a goal, both theoretical and empirical. Many interventions aim to change behaviour. Behaviour is observable and, in this sense, an evaluation could limit itself to using only measures of individual behaviour or its outcomes. But there is a much deeper and broader need. To produce evaluations that are useful to policymaking, we need to understand how individuals behave, and therefore we need appropriate measures of the determinants of behaviour.

Many of the models of individual behaviour we work with, particularly in economics, share a common theoretical framework that is based on some sort of individual rationality or near-rationality: observed actions (i.e. behaviour) are likely to be the product of some subjective valuation of individual gains from that action relative to another action, given certain constraints. In other words, individuals respond to policy interventions (and other determinants of their environment) to achieve certain goals. While at first glance this approach might seem restrictive, these models can be very flexible and include a number of elements that allow for a very nuanced and sophisticated view of individual behaviour. For instance, one can construct and use models where choices are taken on the basis of limited information or distorted beliefs or preferences that incorporate present biases, altruism, and the effect of social norms. However, some version of a structure based on some sort of individual optimisation is key, if evaluators want to identify the causal links that extend from the intervention to the observed outcomes, which, in turn, is crucial to the design of policy

interventions. This structure is also key to the extrapolation of results obtained in a given evaluation (possibly to different contexts) and to achieving improvements in the policies being evaluated.

This set of considerations is even more salient and relevant when policymakers are concerned with the scaling up of an intervention that has been evaluated in a relatively small trial. In that case, the evaluation should consider two important issues. First, the challenge of implementing a certain intervention at scale can be substantial. It is not clear that it is easy to maintain appropriate intervention coverage, and fidelity to the original design. Therefore, a useful evaluation should measure the variables that make an intervention work and, in any case, the availability of the human and other resources involved in a scaling up. An ingredient that is often important for scaling up an intervention is the participation and ownership of the recipients of the intervention and their communities. Second, evaluation at scale should take into account the impacts on other aspect of the economy and social structure (often referred to as general equilibrium effects) that these interventions might have when deployed beyond the realm of a small intervention. It is therefore necessary to model and understand not only individual behaviour but also the interactions among individuals. This brings to the fore the importance of measures (of interactions, of the functioning of markets, of social norms, of existing institutional details and their effects) that might be key for this type of analysis. A good example of the use of evidence from a small randomised controlled trial (RCT) to predict impacts at scale, taking into account general equilibrium effects, is the recent paper by Allende, Gallego, and Neilson (2019), who embed the evaluation of an information intervention among Chilean parents, which shows good impacts with an RCT, within a model that explicitly considers parental choices and the possible reactions of schools to a change in enrolment demand.

Theoretically, within the models we are advocating, individuals choose certain actions based on a variety of factors, ranging from the resources available to them, to the markets they have access to, to their preferences for different outcomes, to their perceptions of the effects of certain actions and their subjective expectations regarding other actions. While some of these factors are directly observable (though not necessarily always in a straightforward fashion), others are not. The latter include what people expect is going to happen (and their confidence in these expectations), people's wishes regarding what is going to happen, people's perceptions of the effects of their actions, the quality of information they possess, and people’s tastes and preferences, and how they are affected by social norms.

Often, when modelling individual behaviour, the lack of appropriate data is overcome by strong modelling assumptions, which permit different types of shortcuts. For instance, we know that many important decisions (on saving, investment, school choices and so on) depend on individual expectations of future outcomes, such as future incomes or the returns to specific investments. These expectations are personal and subjective and can differ from

reality or objective facts. If one does not have data on subjective expectations,[1] a common practice has been to assume rational expectations: that is, each individual is assumed to use efficiently the information available to them. Analogously, in the absence of measurements of individual beliefs, it is common to assume that choices are informed by complete knowledge of the process that generates future variables. In this sense, choices are the result of a calculated balance of costs and benefits and are never 'mistaken'. The need for such strong assumptions effectively makes it clear that without appropriate data and measurement it is not possible to distinguish between individual preferences and individual beliefs. [2] Do certain individuals invest a lot in a certain activity because they believe that activity has a high return or because they have a particular taste (preference) for the expected outcome that such an activity might achieve or because the utility cost of that activity is low for them? Differentiating between these alternative explanations of behaviour can be very important for policy design. A simple example in child development can make this point clear: do parents from disadvantaged backgrounds spend little time stimulating their children because they do not have a very strong interest in their development or because they do not believe such activities are useful or because it is too costly for them to do so? An answer to this question would determine what type of intervention policymakers might want to develop, and it is only possible to obtain an answer to the question by using appropriate measures.

The availability of appropriate data can avoid strong and arbitrary assumptions. In other words, the availability of data and appropriate measurements is linked to the type of models that one can use. At the same time, the development of flexible models of individual behaviour should inform the measurement strategies that researchers and evaluators, in particular, use.

Interpersonal differences in the constraints and the factors that influence perceived gains from certain actions explain a great deal of differences in behavioural outcomes. In the context of evaluation, these elements can be particularly important in seeking to understand how a certain intervention yields certain results. If the scope of an evaluation is not simply to estimate the impact an intervention has on a given outcome of interest in a given context but rather to understand how a certain outcome is originated by behavioural change (or not), it is important to measure mediating factors that determine certain impacts and, importantly, to identify the causal links that exist between drivers of behaviour and the final outcomes. While the recognition of the importance of improving and developing theories of change, there is not always consensus across disciplines on the appropriate method to perform a proper mediation analysis Thinking about a conceptual framework that links human decision-making processes and measurement can provide a specific and useful perspective that stresses the

[1] We discuss subjective expectations and their measurement in Section 4.
[2] This point is raised in a recent paper by Caplin (2021).

importance of identifying, using different techniques and measurement tools, the causal links that are relevant for a coherent use of – and possible improvements of – a theory of change.

In the rest of this paper, we first (in section 3) discuss 'what to measure'. What are the variables of interest to an evaluator? We then move on (in Section 4) to a discussion of 'how to measure' these variables. This is an important – and yet understudied – topic that should receive much more attention. In Section 5, we discuss how to construct new measures, meaning measures of theoretical constructs that have not been formally and explicitly measured before. Finally, Section 6 concludes the paper.

This structure should indicate the main message of the paper: measurement is central to evaluation in many ways. First, measurements methods should be used to address a variety of issues (ranging from comparability to context specific validity) to use effectively existing measures. Second, measurement methods should be used to develop new tools. Such tools can also be key to identifying the causal links that are central to the understanding of the mechanisms that generate the observed impacts of policy interventions. This message is particularly relevant for economists, as a number of measurement techniques and evaluation approaches that we would be advocating have been used in other disciplines – both the use of certain measurement tools and the use of 'theories of change' to interpret evaluation results – but they have not been used much in economics. This is not to say that economists should embrace completely the approaches followed by other disciplines. However, they can contribute to meeting the need to identify the causal links that make the theory of change relevant. More importantly, the adoption of (properly validated) innovative measurement tools can make the identification of structural causal links that inform behaviour easier.

3. What to measure?

The short but not straightforward answer to this question is: what is relevant. Based on the discussion above, one could divide the variables to be measured into two sets: *outcome variables*, which are of direct interest to the researcher or evaluator, and *environmental variables*, which are drivers of the behaviour of the subjects of the intervention to be evaluated. The former might be, for example, the set of variables that a given intervention wants to influence or affect. The latter are variables that might help to identify the mechanisms that generate the ultimate results of the intervention. What variables get included in the two sets depends on many factors, ranging from the type of intervention one is evaluating, to what the outcome variables of interest are, to the methods of evaluation that are feasible. In this section, we discuss both the measurement of the outcomes of interest and of additional variables that can be useful in understanding the impacts of an intervention.

3.1 Measuring outcomes

In many cases, the measurement of outcomes is reasonably straightforward. In some situations, however, certain outcomes are difficult to measure, especially in the context of developing countries. In such a situation the best strategy and approach is to explicitly recognise the presence of measurement errors and to devise the data collection in such a way as to be able to minimise their effects. As a perfect measure does not exist, it might be better to have two imprecise measures of the same variables than one more precise one, provided the measurement errors affecting the two variables are independent of each other. In such a case, one could use one variable as an instrument for the second, or, more efficiently, embed them in a measurement system that could provide efficient estimates of the (latent) variables of interest.

A related issue, which is more specific to developing country contexts, is the use of standardised tests that typically put a large number of different items through a scoring algorithm. The typical example here is the use of tests of child development which have become the standard in the international literature. In most cases, the algorithms that provide the estimated scores of child development from the available items were constructed many years ago using samples typically from WEIRD (Western, educated, industrialised, rich, and democratic) countries. It is not clear that the same algorithms, which effectively determine the weights received by different items, would be valid and efficient in a developing country context.

An alternative strategy would be to use the individual items to construct new scoring algorithms that, using the available individual measures, make an efficient use of them, taking into account that certain items might be more or less informative, depending on the context

in which they are collected. A simple example can help to make the point here. Many tests of child development for the early years measure the ability of the child to recognise certain pictures and to name the object they represent. The ability of a child to recognise the picture of a boat would reflect a different level of development depending on whether the child lives in a port town or in the middle of a desert. Constructing a new scoring algorithm is now feasible and easy even with the most standard software used in evaluation.

Yet another strategy, particularly relevant for developing countries, is to construct new measures. A large effort is underway within the Global Scale for Early Development project to construct new measures of child development that are more relevant in certain contexts, which are very different from those in which the original measures were developed and, importantly, that are easier to collect at scale in developing contexts.[3]

3.2 Measuring to understand mechanisms

The second message we want to impart is about the importance of measuring not exclusively the outcome of interest but rather a set of variables that can be used to understand how certain interventions achieve the observed outcomes. Our argument can be conveyed effectively through a number of examples.

Suppose the purpose of the evaluation is to establish the effect of a deworming drug on the health status of children and on children's educational attainment and development.[4] And suppose that the researcher has performed an RCT where a deworming intervention is implemented in a randomly selected group of villages, chosen from among a wider set of villages. In these villages a number of children and their families are recruited in the evaluation sample.[5] If this was a medical experiment, possibly conducted in a lab, one would measure the health outcomes of interest, for instance the effectiveness of the deworming procedure, and possibly some nutritional outcomes, such as height, weight, anaemia, and so on, of the children in the treated villages, and one would compare these to outcomes measures for the children in the control villages. The randomisation at the village level and the complete coverage of a village or school is important in this context because of the potential (negative) externalities that a partial treatment might have, leading to re-infection and the like. Moreover, in addition to the final outcome, in this context it might be important to measure adherence to and compliance with the intervention protocol at the school level and for groups of children. What we want to stress is that the individual-level outcome is not

[3] See https://earlychildhoodmatters.online/2019/the-global-scale-for-early-development-gsed/.

[4] This example is inspired by Kremer and Miguel (2004).

[5] In these examples we are not considering ethical considerations nor a discussion of the appropriate protocols that should be used in the data collection and in the recruitment of the evaluation sample. This is not because these issues are unimportant, but because, in the present context, we want to convey a specific message about what variables should be measured.

the only relevant variable, because the outcome will be affected by important interactions and externalities. These features of the experiment environment can and should be measured and the resulting variables should be used to model the effect of the intervention. The results of such analysis could give important information on, for instance, the relevance of reinforcement campaigns or the effects of different dosages of the intervention.

Consider then an intervention that aims to improve the position of women in the family. An intervention of this kind, which aimed to increase the control of financial resources given to women, is evaluated in Field *et al*. (2021). In that paper, in addition to establishing the impact of the intervention on certain variables (such as financial activity and labour supply), the authors discuss the mechanisms that could have led to such impacts, which might be in contradiction of standard models of intrahousehold allocation that are often used in economics. A model of household choices when more than one decision maker is present is the so-called collective model, proposed by Chiappori (1988). Within such a model, increasing the bargaining power and control of a spouse (in this case the wife) should lead to higher consumption (of commodities but also leisure time) by that spouse. This candidate mechanism contradicts the finding of the evaluation exercise which documents an increase in female work hours and labour force participation. One possible explanation of these results is the influence that social norms might have on individual preferences: when wives have less control, they might be constrained in their labour supply choices by social norms and their husbands' views, influences that might be reduced by a shift in control. A potential approach to this problem, then, is the measurement of such social norms, which, as we discuss in Section 4, might be difficult.

As an additional example, consider the evaluation of a stimulation intervention targeted at young children in rural towns in Colombia, which consisted of weekly home visits to the houses of young children to improve parenting practices. Attanasio *et al*. (2014), using a clustered RCT, showed the intervention had an impact of 0.26 standard deviations (SDs) on children's cognitive development. An important question then is what drove those results. To answer that question Attanasio *et al*. (2020) estimate a production function model in which child development depends, among other things, on the initial value of development and parental investment. As the intervention induced a remarkable increase in parental investment, the question is whether the impact of the intervention worked through such an increase or was induced by other factors, such as the weekly contact with the person conducting the visit. To answer this question, it is key to establish the *causal link* between parental investment and child development, which is not an easy task, as parents might react to a variety of factors linked to child development. Attanasio *et al.* (2020) use an instrumental variable approach to establish such a causal link: that is, they model parental investment as being determined by a set of variables, including some that, plausibly, do not have a direct effect on child development. It is important to stress that, in this context, the randomisation

of the intervention cannot be used as a valid instrument (despite being controlled by the researchers) because the question that is posed is whether the intervention worked *only* through parental investment or whether it had a direct effect (possibly through the weekly exposure to the visitor and the activities she was engaging in, therefore violating the exclusion restriction). Attanasio *et al*. (2020) use variation in the prices of toys, books, and food across the towns where the programme was implemented, and the exposure to violence experienced by the children's mothers when they were adolescents, as 'instruments' – that is, variables that can affect parental investment (which they do) without having a direct impact on children development (which has to be assumed). Having identified the causal link the authors then show that most of the impact of the intervention can be explained by an increase in parental investment. In the present context it is important to stress that the availability of data on past exposure to violence and on prices was crucial to the identification of the structural model used for the mediation analysis.

The next question considered in the same evaluation is why parents increased their investment. One possibility is that the intervention changed parental beliefs about the process of child development and about the usefulness of parental investment in this process. In another, Attanasio *et al*. (2019) elicited, in a second follow-up, information about individual beliefs of the process of child development and of the importance of parental stimulation. In that paper, the authors show that (i) in the population being studied, parents seemed to underestimate the productivity of parental investment, especially for children with high starting level of development; and (ii) that the perceived productivity of investment was predictive of actual investment.

Here it is clear that, in some contexts, to understand how an intervention works it may be important to develop complex and novel measures that have to be devised and implemented carefully. Attanasio *et al*. (2019) describe the specific methods they used to elicit parental beliefs.

This discussion, and in particular the point about the impact of individual beliefs on individual choices, makes it clear that without direct measures of beliefs it is not possible to disentangle the extent to which observed individual choices are driven by preferences or by beliefs – a point that has been made forcefully by Caplin (2021).

Another related example is that considered by Calvi and Keskar (2021), who study the effect of reducing the practice of dowries in India. In their paper, they use variation at the state level to point out that such a well-intended policy might actually have negative impacts on women's wellbeing, changing the nature of interactions within the marriage and even the type of sorting within the marriage market. While the paper uses a clever identification strategy (using gold prices in the year of marriage and variations in the Indian anti-dowry law in the 1980s),

their case could be made stronger by explicit measures of bargaining power within the marriage.

A final example, in which such novel measures of bargaining power within the marriage were constructed, relates to the evaluation of the impact a certain intervention might have had on the position of women within the household. In Macedonia, a conditional cash transfer programme was randomly allocated to women in certain towns and to men in others. Almas *et al*. (2018) construct direct measures of relative bargaining power within couples to measure the impact of this programme and show that indeed the targeting of women had a significant impact in this dimension.

Richer theoretical frameworks require richer measurements. In practice, to bring a realistic theoretical framework to data so that it has empirical bite, one needs a measurement system. In what follows, we discuss a number of, in our opinion, noteworthy points about measurement, their construction, and their use, with specific reference to evaluation work. We start with four notable considerations regarding 'how to measure' constructs of theoretical models that are not directly observable. We then discuss some issues related to the challenges encountered in the construction of new measures and their validation.

4. How to use measures and how to measure impacts

Depending on the outcome and the mediating factors one wants to measure, the methods used to obtain measures that can capture such outcomes and mediating factors can differ. A very important issue here is the metric that is used to measure outcomes, in order to compare them across treatment conditions and across contexts. This section discusses a number of overarching considerations that are important in all evaluation studies. These include: a) explicitly recognising measurement errors; b) measuring impact sizes; c) ensuring interpersonal comparability of measurements; and d) measuring behaviour through stated versus revealed choices.

a) **Explicitly recognising measurement error**

Measurement error is a pervasive challenge for most applied work. It can arise from a variety of sources, including (i) missing data and non-responses, (ii) difficulty in measurement, and (iii) misreporting. More generally, the variables (or latent factors) one is interested in are not necessarily equivalent to the available measures.

The consequence of measurement errors in explanatory variables is a bias in the estimated regression coefficients. If a measurement error is of the classical type (the measurement error is uncorrelated with the true value and is additive), one typically gets what is known as an 'attenuation bias': that is, the size of the relevant coefficient is reduced in absolute value. In the case of a non-classical measurement error (when the measurement error might be correlated with the true variable and enters non-linearly) it is impossible to assess the size of the induced bias.

The consequences of measurement error in the dependent variable are larger standard errors in the coefficient estimates. In some sense, this consequence is less severe because causal inference is still possible, although predictions and policy simulations are noisier.

In the case of a classical measurement error, one can interpret measurement error in the explanatory variables (typically known as the errors-in-variables problem) as a case of endogeneity in the explanatory variables: that is, measurement error induces some correlation between the unexplained variation of the left-hand-side variable (what is usually referred to as 'the residual' of the equation under study) and the variable for which one wants to isolate the impact or its measures (the observed value).

More specifically, let $x = x^* + \eta$ where x^* is the true value of the variable and x is the measure observed empirically: it contains the true value and measurement error η. The

'endogeneity' bias introduced in this situation in a regression equation relating the variable *x** to a certain variable *y,* is easily seen. Suppose that the relation of interest is:

$$y = \beta x^* + u \qquad (1)$$

where x* is unobservable. Equation (1) can be converted in terms of the observable *x* as follows:

$$y = \beta x + \eta + u = \beta x + e \qquad (2)$$

where $e = \eta + u$. Obviously, the residual of equation (2) is correlated with *x.*

In linear models, one solution is to adopt a second measurement of the true but error-prone variable, say *z=x*+υ.* An important condition that must hold in order for a second measurement to help out is that the measurement error *υ* is uncorrelated to the first measurement error, η, and both these errors are uncorrelated with the outcome variable. The second measurement is used in the same way as an instrumental variable: the instrument is correlated with the mis-measured variable but not correlated with the measurement error. The treatment of measurement errors in non-linear models requires further assumptions. For a discussion of measurement errors in non-linear models see Chen *et al*. (2011).[6]

An alternative that is often used in applied work to reduce measurement errors is to average two (or more) measurements and to use the combined average measure in estimation: assuming the two measurement errors, η and *υ,* are uncorrelated, as with Instrumental Variables (IV), averaging will reduce the combined measurement error.

While the IV approach and the averaging approach can be useful, they are not necessarily the most efficient methods. They require multiple measurements but only use the correlation among them (the IV approach) or aggregate them (averaging), assuming certain weights, therefore reducing the importance of certain measures that could be very informative. A more efficient alternative is to develop an explicit measurement system: this requires identifying the parameters that link underlying latent (unobserved) factors to their measures (observed indicators). Several examples of such systems exist (e.g. the MIMIC model, factor models, item response theory). Precise specification of the measurement system and the assumptions that will be required to use the observable indicators should inform the way data are collected.

In empirical applications with primary data collection some design features can minimise the influence of measurement error. As mentioned above, if one wants to use two or more measures to address the measurement error issue, a key assumption is the fact that the measurement errors of these variables are independent. This could be

[6] For measurement error adjustment related to networks see Advani and Malde (2018).

achieved when collecting data: for instance, by assigning different enumerators to collect different measures on the same subjects. While such a strategy might not be feasible in certain contexts (especially for small evaluations and surveys), it is feasible for larger surveys. More generally, there are a number of creative ways, during the data collection, to make sure that certain properties of the measurement system, which would warrant the use of certain techniques, are satisfied. The message here is that an explicit consideration of the pervasiveness of measurement error can be useful in the design of survey methods.

Other design elements can feature directly in the data elicitation procedures. This is particularly relevant for data collection where self-reporting bias is expected. This type of bias should be expected whenever one is interested in measuring behaviour or opinions which might be perceived by respondents as 'inappropriate'. Respondents may choose to misrepresent their reporting of their attitudes because of concerns about their social image in the eyes of the researcher, themselves, or others (a phenomenon called social desirability bias), or they may have beliefs about what is expected from them by the researcher and may modify their behaviour accordingly (a phenomenon called 'demand effects').

These types of effects bias the data in a way that is difficult if not impossible to adjust for econometrically, after data collection. To minimise social desirability bias and demand effects, using appropriate experimental designs can be a useful way to elicit truthful self-reported behaviour. However, in this case (as in any situation where new measures are piloted), validation exercises are particularly important. This type of design has been applied, for example, in studies of lying behaviour (Fischbacher and Foellmi-Heusi (2013); see Box 1), labour discrimination (e.g. Eriksson and Rooth, 2014), attitudes towards immigrants (Hainmueller *et al*., 2015), and stigmatising behaviour relating to Covid-19 (Cavatorta, 2021). For a discussion of demand effects in experiments, see Zizzo (2010).

Box 1

Lying falls into the class of behaviours that are socially framed as inappropriate or morally 'wrong'. For this reason, self-reported answers related to one's lying habits are likely to be misreported (arguably, underreported). Respondents are likely to change their behaviour (or reported behaviour) according to what constitutes 'appropriate' behaviour (in some sense, what is 'demanded' from them, hence the term 'demand effects'). Fischbacher and Foellmi-Heusi (2013) design a measure to detect honesty and lying that minimises demand effects. Their measure is simple: participants receive a die and are asked to roll it privately and report the number. The payoff is linked to the reported number, which is done privately and anonymously. Since the true probability of each rolled number (in a sufficiently large sample) is known (it is ⅙), the authors can compare the empirical frequency of each reported rolled number with the expected frequency. They find that about 20% of participants in their experiment lie to the fullest extent, while 39% are fully honest. This design allows for the detection of lying behaviour at the group level; exactly because the design minimises demand effects, it cannot detect lying behaviour at the individual level.

b) **Stated versus revealed choices**

Measurement of individual behaviour can be divided into two categories: observed behaviour (or revealed choices) and (self-) reported behaviour in hypothetical situations, (or stated choices). Traditionally, in economics, observing behaviour directly has been preferred as a measurement in evaluations because it minimises different types of self-reporting bias. However, in many contexts observing behaviour can be limited or impractical. Data on what people actually do (e.g. administrative data) may not be available, or accessible, or the logistics involved in having observants (e.g. class observations) may not be feasible. Recording the behaviour of subjects, either in their natural environment or in purposefully designed sessions, and later coding the behaviour, is an approach that is increasingly used.[7]

Eliciting self-reporting of behaviour is a common way to measure behaviour, particularly in surveys. Yet measuring self-reported behaviour is not always feasible or appropriate, and may be prone to misreporting, as discussed in the previous section. Illegal, anti-social, and more generally sensitive behaviour are all examples of areas where self-reporting of behaviour is prone to significant bias. Demand effects and social desirability bias pose a threat to external validity because if the experimenter were absent the subjects would adopt different behaviours. While these biases are concerns in measurement in general, they can be particularly important for the evaluation of interventions such as educational, crime prevention, and social attitude change interventions. In extreme cases, one can imagine that repeatedly emphasising the benefits of a specific behaviour during a training and then asking subjects to self-report

[7] Such an approach has been used, for instance, in classroom settings to record the interactions between teachers and pupils. Analogously, in some settings interactions between parents and children are recorded.

how much they would adopt that behaviour can lead to significant demand effects among the treated group, obscuring the interpretation of the real impact of the training.

In many instances, behavioural outcomes that are of interest to social scientists are context-dependent, in the sense that they depend on individual circumstances at a specific time. Think about aggressive behaviour in social interactions. Observing 'aggression' depends on individual incentives, individual constraints, and the presence of a situation where aggression can – but need not – arise. Direct observations of aggressive behaviour are likely to be difficult and hardly interpersonally comparable because people are unlikely to find themselves in the same situation. In a survey on aggressive attitudes, people would be reluctant to openly admit they engage in, or have engaged in, aggressive behaviour and thus self-reported behaviour is likely to be mis-measured.

In these circumstances, experimental interactive tasks, which are widely used in experimental economics, can be a useful tool. Interactive (sometimes gamified) tasks place participants in common decision situations that, albeit abstract, resemble interpersonal interactions in real life. Since participants' constraints, incentives, and information are under the control of the researchers, these tasks offer a way to provide interpersonally comparable measures on individual behavioural inclinations. Furthermore, as participants' behaviour in these tasks is directly observed, this approach can reduce the problem of self-reporting bias. There are an increasing number of examples of experimental tasks that are conducted in the field, embedded within larger survey modules. Box 2 presents an application measuring a type of anti-social behaviour: vindictive behaviour.

Box 2

Cavatorta *et al.* (2020) evaluate whether exposure to violent interactions influences behaviour towards classmates, and potentially begets more violence. They look at the case of Palestinian school children who are obliged to cross Israeli military checkpoints on a daily basis to go to school. Such checkpoints can be shown to be locations of frequent violent interactions. The study compares retaliatory behaviour in groups of children going to the same school, some of whom had an obligation to cross checkpoints on their way to school (and some of whom did not). The authors use an interactive game between pairs of participants who play in turn. There are clear and simple rules of the game. Each player receives an endowment: no aggressive action and no retaliation is the socially optimal choice for both players; but the first mover can perform an 'aggressive' action towards the co-player by taking away some of the co-player's endowment. In the following turn, the co-player can decide whether or not to retaliate against the first mover's action. Retaliation need not occur in the game but if it does the second player's behaviour captures a marker for vindictive behaviour, which is comparable across participants. Every player is in the same situation and faces the same incentives and constraints. The directly observed behaviour can be correlated with (self-reported) stated choices or administrative data (e.g. school disciplinary action). The authors find that youths that were more exposed to violence were significantly more likely to display retaliatory behaviour against their peers.

However, interactive tasks involving participants are not without limitations. One concern regards the abstract nature of the interactions. Against this backdrop, recent efforts have been made to implement gamified tasks in a truly interactive manner, and to make them more realistic: for example, by using video game-like tools. These tools are particularly promising in efforts to measure attitudes and inclinations among youths. Interactive tasks and games do not rule out demand effects, but they can significantly minimise them (e.g. using design features, careful language in instructions, and anonymity between participants and between participants and experimenters). A final note relates to the external validity of these measures. Interactive games have traditionally been developed in laboratory settings or purposefully designed RCT. Relatively little evidence has been produced on how much variation of behaviour outside the experiments these measures are able to explain and whether, therefore, they can be used as predictors of behaviour in the realistic circumstances that they try to mimic. Sometimes behaviour in lab environments is at odds with behaviour in the field (List, 2006). While measures based on experimental tasks score high in terms of internal validity, their external validity remains debated and this is an active area of research (Levitt and List, 2007; Harrison and List, 2004; Galizzi and Navarro-Martinez, 2019).

Another useful approach to measurement on which a large literature exists is that of vignettes, which are widely used. However, this approach is subject to a variety of problems and issues, especially in terms of interpersonal comparability. A large

literature on these issues exists (see, for instance, King and Wand (2007) and Hopkins and King (2010)).

Our discussion so far has focused on the fact that some behaviours or drivers of behaviours can be hard to observe, or the fact that participants might find it difficult to report their choices truthfully. A related dimension of measuring 'stated' choice is the elicitation of hypothetical choices in a set of circumstances presented by the enumerators. This technique, which is widely used in marketing and other disciplines, was tentatively used in economics in the 1950s and 1960s. An example of the use of hypothetical choices is Juster and Shay (1964) and Juster (1964). Thereafter, these techniques were rarely used in economics for a number of decades (but see Erdem *et al*. (2005)). However, more recently a number of studies have started to use this approach for the explicit task of estimating some of the parameters of structural models. Recent examples include Americks *et al*. (2020), Caplin (2021), Bernheim *et al*. (2022), and the discussion in Almas *et al*. (2022). The motivation behind using this approach is very powerful. Often, in the empirical study of structural models, the identification of the causal link between two variables is made difficult by the fact that the variation in the two variables is driven by potentially correlated and unobserved factors, often related to individual behaviour. By introducing hypothetical scenarios and eliciting choices under those scenarios, researchers can control the variability of one of the variables of interest, and therefore deduce the causal links of interest. Of course, the elicitation of such stated preferences and hypothetical choices is not a trivial exercise and it is fraught with a variety of challenges, some of which we have mentioned above. For this reason, validation exercises, especially analysing data on stated preferences jointly with data on revealed preferences, can be particularly useful.

c) **Impact sizes and scaling**

Using evaluation to understand 'what works' implies the need to estimate the size of a treatment effect. Different metrics can be used to evaluate the treatment efficacy in a *given* context, but metrics are to some extent inherently context-dependent (think about how the average human height varies across ethnicities). If the objective of evaluation is to make objective comparisons *across* contexts, the effect size needs to be dimensionless. This is often achieved by standardisation.

There are various ways to standardise the difference between control and treatment groups, but each way comes with its own pros and cons. A common way to standardise treatment effects is to use the ratio of the difference between treatment and control groups' average outcomes (i.e. the treatment effect coefficient in a regression framework) to the standard deviation of the control group's outcome (or baseline period). The typical interpretation is 'how much of a standard deviation difference compared to the control group the treatment generates'. However, while such an approach can provide a meaningful metric in some contexts, it can be very misleading

in others. If the standard deviation of the control group reflects some 'natural' variation in the outcome of interest, such an approach could make sense. However, the same impact could look large if the policy is targeted at a very homogenous population and very small if it is implemented in a very diverse population. So what alternatives are available?

One alternative way of interpreting the size of an effect relies on 'contextualising' it. A first way to contextualise the impact of an intervention is to compare the value of the treatment effects obtained in one context with known magnitudes from other studies or known values from meta-analyses. A second, and maybe more attractive, alternative is to start comparing the population being studied (typically individuals or households in a disadvantaged context) to a representative sample from the same country or region but including individuals from different backgrounds. This makes it possible to position the target population within the overall population and to measure the size of the impact in terms of movements within that population. For instance, suppose that the target population has, before the intervention, an average outcome that is similar to that of the bottom 20% of the representative sample. One can then measure the impact of the intervention as a fraction of, say, the difference between the 80th and 20th percentile in the representative sample (see Box 3 for one example). This type of metric indicates the extent to which an intervention is able to remediate initial inequalities and to help the disadvantaged group 'catch up' with those who start from a less disadvantaged position.

Box 3

Many policy interventions have a remedial goal: levelling up beneficiaries' outcomes with those of others who start in a less disadvantaged position. Heckman et al. (2014) assess the impact of an early childhood development intervention on later-in-life earnings. In addition to comparing the earnings of a group of stunted children receiving the intervention (treated group) with a group of stunted children not receiving the intervention (control group), as in traditional evaluation designs, they compare the earnings of stunted children receiving the intervention with the earnings of non-stunted children. If treated stunted children 'catch up' with initially better-off children, this indicates that the intervention is able to remediate initial inequalities. It also provides a useful comparative estimate that can be used to evaluate the opportunity cost of different policies that aim to produce improvements in the same outcomes.

A third alternative is to express the size of a treatment's impact in terms of its economic returns. This is often achieved by adopting a monetary metric. This approach rests on the availability of data on a monetary reward and being able to relate the monetary and non-monetary effects of an intervention. While such an approach is particularly attractive, as the monetisation of the benefits of an intervention can then be easily compared to its costs, the exercise can be extremely difficult. Consider, for instance, the evaluation of an intervention aimed at improving early childhood development. While we now know that development in the early years is extremely predictive of long-term

outcomes, trying to relate, explicitly, impacts on different measures of early development (possibly in different dimensions, including cognitive and socioemotional skills) to adult outcomes, such as earnings, can be very difficult. Even ignoring potential general equilibrium effects that may change the returns on certain skills in the labour market and technical changes that can induce additional changes, it is difficult to relate early and late outcomes for older cohorts because of the scarce availability of data. The best strategy in such a context is to rely on previous studies and to make clear that some conclusions might be tentative.

Another related issue in this context is the fact that earnings might not be the only outcome of interest or of relevance in the long run. One could therefore try to incorporate other aspects, such as health outcomes, life satisfaction, or criminal behaviour. These issues are amply discussed in the literature on willingness-to-pay valuations in cost–benefit analyses.

d) **Interpersonal comparability of measurements**

When measuring constructs for quantitative analysis, the underlying assumption is that measures are interpersonally comparable so that summary statistics can be reasonably computed and are meaningful.[8] We discussed earlier the importance of data on expectations. Expectations feature in many behavioural models of decision-making. Expectations of benefits from an action, such as adopting a particular behaviour or a treatment, can explain behavioural change and treatment compliance. As mentioned earlier, expectations can be an important mechanism for the success of an intervention. For example, think about child stimulation interventions: even though the direct intervention applies to the child, whenever mothers of recipient children develop a stronger belief in the importance of child stimulation, they tend to adopt behaviours which reinforce the positive impacts of the intervention on the child.

The traditional way (though it is still widely applied) to measure beliefs about gains or expectations of things happening is to use qualitative categorical responses concerning whether a certain event is deemed ‘unlikely’, ‘somewhat likely’, ‘likely’, or ‘very likely’. While this type of data is undoubtedly interesting information, this type of measurement illustrates the problem of interpersonal comparability. The definition of what is 'likely' is not identical across people: some people might think of 'likely' as a chance of eight in 10 and some people may think it is a chance of five in 10. In a recent study, Wintle *et al*. (2019, Figure 5) illustrate how wide the different interpretations of target verbal likelihood phrases such as ‘very likely’ can be. When participants were asked to translate

[8] This section refers to interpersonal comparability from a measurement point of view, intended in an empirical and data collection sense. On the interpersonal comparability issues, see also King *et al.* (2004). A broader and more sophisticated debate relates to interpersonal comparisons of utility, which have had an important role in welfare theory and social choice theory, and the related debate about the comparability of happiness and well-being. We refer here to Barbera *et al*. (2004).

likelihood phrases from words into numerical estimates, 'Very unlikely' was translated into probability figures ranging from 0% to just over 40%, while 'very likely' was translated into figures from around 65% to 100%. For this reason, responses to categorical questions of this kind score very poorly on interpersonal comparability.

To elicit the perceived likelihood of certain events, probabilistic assessments present several improvements over traditional qualitative categorical responses as regards interpersonal comparability. Probabilistic assessments elicit the respondent's perceived percentage chance of well-defined events occurring (what is the chance of event E happening?). There are several examples of the use of these measurements in relation to education (Attanasio and Kaufman, 2017; Attanasio, 2009; Wiswall and Zafar, 2021), job expectations (Manski, 2004), consumer confidence (Dominitz and Manski, 2004), and political behaviour (Cavatorta and Groom, 2020). Probabilistic assessments have the advantage of being interpersonally comparable since probabilities have the same meaning across different people.

Probabilistic assessments generally require that one estimate (a point estimate) be made by respondents, although events may be highly uncertain, and people differ in their tendency to report the best estimate, the worst estimate, or anything in the middle. Take for example the case of income expectations. Respondent estimates of how much their income is going to be in a given period in the future are inherently uncertain. There is a distribution of possible outcomes: some figures might be more likely than others. In a survey, some respondents may be reporting the minimum value expected, some the maximum value, and some may be reporting the mean value or the mode (the more likely for them) or the value that is more salient to them. For the researcher, the inability to distinguish between different interpretations of the same question is problematic. This can be partly mitigated by careful wording. For example, questions may directly ask for the 'most likely probability that event E will occur' or the 'highest/lowest probability that event E will occur'.

A separate issue is that the uncertainty surrounding the likelihood of a specific event can be of direct interest. In the case of income expectations, the expected variance in one's income may be the factor driving behaviour, rather than the expected average level. Think about the demand for insurance against risk: risk-averse subjects would tend to reduce the variance of expected risks. Box 4 describes a method that can be used to elicit the entire distribution of income expectations. This method has been extensively used in developed and developing countries, though it remains a relatively involved and time-consuming measure.

Box 4

This method is particularly suitable when the objective is to approximate the respondent's subjective probability distribution for a variable of interest. For example, eliciting the respondent's expected income distribution would work as follows. The respondent is initially asked the minimum and maximum value of their income at a given point in the future (e.g. 12 months from now). These values constitute the support of the distribution. This support is split into specific interval thresholds (measured in amounts of income), a(1)<a(2)<a(3), and the respondent is asked questions such as 'What is the percentage chance that your income is less than or equal to a(j)?' The combinations of a(j) and the corresponding reported probability are used to make inferences about the respondent's subjective cumulative distribution, from which the expected mean and expected variance can be derived by fitting some distribution on the data. A typical comprehension check is that the reported probability increases monotonically, since the probability that income is less than the reported maximum value should be 1. This method has been implemented in several settings and using different interview modes, in person (in particular in developing countries, see Attanasio, 2009) and by telephone (Dominitz, 2001; Cavatorta and Groom, 2020).

An inevitable issue in eliciting the perceived likelihood of events or subjective beliefs (as well as other measurements) is the level of their accuracy. There is no guarantee that respondents will state the true subjective probability. Respondents may choose to distort their answers in order to rationalise past actions, in other words demonstrating to the researcher or themselves that they made the right decision (i.e. predicting an increase in house prices if one has just bought a house), or they may report their wishing-thinking probability. Respondents may report salient probabilities (e.g. 0.5 or 1) as a cognitive 'short cut', or they may try to preserve a positive self-image (i.e. what is the chance that you will donate to a homeless person?). This concern has led to the development of several mechanisms to incentivise respondents' trust-telling when eliciting their beliefs. The simplest way to do this is to elicit a frequency guess regarding how many instances out of N instances results in a specific outcome: one then compares the frequency guess with an objective realisation (or an appropriate statistic of the realisation) and rewards respondents who are correct or approximately correct (see Box 5 for an example). More complex mechanisms exist. These include very complex mechanisms that are applicable to situations in which the truth is not verifiable (e.g. the Bayesian Truth Serum discussed in Prelec (2004) or the choice-matching mechanism applied by Cvitanic *et al*. (2019)). For an excellent survey of these mechanisms see Charness *et al*. (2021).

Box 5

A simple way to incentivise truth-telling of beliefs is to reward participants when their guess turns out to be correct empirically. In an evaluation of the effect of exposure to violence on actual behaviour towards others among adolescents, Cavatorta et al. (2021) were interested in measuring the participants' expectations about other people's behaviour. Own behaviour may be driven by whether one expects kindness or unkindness in return. Behaviour was measured using a simple gamified interactive task involving two participants: a first mover could take away tokens from the second participant and the second participant could then decide to retaliate or decide not to. Cavatorta et al. (2021) elicited beliefs by asking 'how many participants in this room who play the role of first mover will take away some tokens from their co-player?'. Participants were rewarded if their answer was factually correct. This is a simple elicitation mechanism that is easy to understand (arguably easier than eliciting a subjective probability) and easy to implement in field studies. From a theoretical perspective, this method elicits the mode of the distribution over all possible empirical frequencies of an outcome. However, this method does not work for binary variables with no repeated draws: for example, 'what is the likelihood that Italy will win the World Cup' requires a probability estimate as an answer.

Another situation in which there are challenges in achieving interpersonal comparability is when respondents are asked to report the perceived benefits (or costs) of specific situations, actions, or policies. In these cases, the challenge to interpersonal comparability comes from different interpretations of the counterfactual situation without that action or policy. To draw interpersonally comparable inferences, it is important to maintain the counterfactual in respondents' constant across respondents. The simplest way to do this is by framing or explicitly outlining the counterfactual situation: 'What would be the percentage chance of event E under policy A, compared to a situation under policy B'. Sometimes the difference in expectations between two or more situations is the measure of interest. Box 6 describes an application by Cavatorta and Groom (2021): the goal of the study is to measure the perceived benefits from peace negotiations relative to a scenario of no negotiations (status quo) between two parties that are in conflict.

Box 6

In a study on conflict resolution, Cavatorta and Groom (2021) illustrate the use of a survey design to measure the (perceived) benefits of engaging in political negotiations. The authors' design aimed to) guarantee that the comparison scenario (i.e. not engaging in negotiation) was well-defined and was the same for every respondent; ii) take into account that peace negotiations can end up in different peace deals (or even the failure of a negotiated agreement) and respondents needed to have in mind comparable potential peace deals. This was achieved by presenting respondents with concise descriptions of a number of possible outcomes of negotiations (e.g. in the authors' case study on the Israeli-Palestinian conflict these were: a pro-Israeli peace deal, a pro-Palestinian peace deal, a balanced peace deal, and a failure of negotiations). Cavatorta and Groom then elicited the probability of a set of benefits conditional on each outcome of the negotiations directly from respondents. The questions were relatively simple and were of the form: 'if this scenario happens, what is the chance that …' or 'If this scenario happens, what do you think the level of X will be? Or how do you think X will change?'. The data can be used to measure the expected returns from peace negotiations: in other words, the expected difference in benefits arising from engaging in peace negotiations compared to not engaging in them (taking into account the set of possible outcomes of negotiation, and the perceived probability of each of these outcomes – in other words computing an expected value). The

Some limitations of probabilistic assessments are worth mentioning. i) These instruments require a basic understanding of likelihoods and thus might not be suitable for very young or illiterate respondents. Visual aids and warm-up questions can be used to facilitate understanding: for example, small coins or stones can be used to illustrate the notion of probability. ii) Subjective probabilities from respondents, even from those who are familiar with the concept of probability, are not guaranteed to sum up to one (an issue referred to as additivity). This can be problematic in estimation. This problem may be mitigated by an appropriate survey design and the use of visual aids or help messages. iii) There are issues relating to the levels of respondents' confidence in the elicited probabilities. This which might be worthy of analysis, as in the literature on limited awareness (e.g. Karni and Vierø, 2015), beliefs ambiguity (e.g. Giustinelli and Pavoni, 2017), and theories of learning under ambiguity (e.g. Epstein and Schneider, 2007). Confidence levels for respondents' assessments of chance can be proxied empirically by asking respondents to indicate the range of probabilities of a given event, by expressing qualitative statements on point estimates, or by asking respondents to assign weights over a range of states (e.g. each possible figure in a probability range deemed possible).

5. How to construct new measures

It is a standard practice in academia to implement existing tests and measures that have been used for a long time. It is common to use scoring mechanisms that have been designed in a different context, which might be different from the context where the intervention to be evaluated is implemented. Such an approach is not necessarily efficient and can lead to serious biases. The construction of new measures, when appropriate, both in terms of the factors one is trying to measure and the tests one uses, is an important endeavour.

From a theoretical point of view, some of the important properties of new measurements include the following: *validity* – the degree to which the measure is recognised to accurately proxied what it is intended to measure (this applies also to non-physical or non-immediately quantifiable constructs, like empathy or depression); *reliability* – the degree to which a measure remains constant when is not expected to change or when the underlying conditions remain; and *variability* – the degree to which the measure distinguishes between interpersonal differences. From a practical point of view, desirable properties include how easy it is to administrate and understand the instrument. Often there are trade-offs across properties, such as validity versus practicalities. These trade-offs affect the degree of measurement error included in any specific measure.

Methods to synthesise information are extremely useful in order to achieve a compromise between validity and practicality. Data reduction techniques like principal component analysis (PCA) – a statistical technique that extracts the linear combination of the data which explains as much as possible of the variation in the original data – are particularly useful to summarise a wide range of indicators on a smaller set of 'components'. These components can be interpreted as latent factors which underlie the indicators (data): different indicators might be strongly correlated with one factor and less with another, and the strength of the correlations will show up in the loading parameters of the PCA. Often the first factor is the one that is of interest. The indicators that load more strongly on the first factor are those that are more informative about it. This strategy has been adopted, for example, to optimise the collection of information that traditionally has required a large and time-consuming battery of questions (see Box 7 for an example relating to child development measures).

Box 7

The Bayles Scales of Infant and Toddler Development is a popular measure of child development. It contains 91 items that are typically asked, one by one, of the child's mother. Such a high number of items limits its use in resource- and time-constrained data collection situations. In a case like this, PCA can help identify the most informative items. Attanasio et al. (2020) collected the Bayles Scale to measure cognitive ability in a sample of Indian toddlers and used PCA to identify the most informative items. The result was a set of 15 items. Using the linear combination of the 15 items as a proxy for the latent factor of cognitive ability yielded a distribution equivalent to that obtained using the entire set of 91 items. The implementation of the shorter questionnaire required approximately a sixth of the traditional time required to implement the full Bayles Scale. This is not to say that the short-version of the Bayles Scale is superior to the long version or vice-versa, but in many settings the 15-item scale can be more practical, if not the only possible avenue.

There are complex constructs, like emotional intelligence, which cannot easily be easily quantified with one indicator or even a few indicators. There may be multiple latent factors underlying a set of indicators. If there is a good theoretical basis for what these factors might be, it is possible to analyse the covariates of these factors. Emotional intelligence is an example of a latent factor that is reflected in a multitude of 'indicators' (these could be the answers to a battery of questions related to emotional intelligence). The Multiple Indicators Multiple Causes (MIMIC) modelling, a special case of Structural Equation Modelling (SEM), offers a statistical approach to summarising the way the latent factor manifests in some 'indicators' and the way in which it can be influenced by certain 'cause' variables (in the sense of covariates). The MIMIC model consists of two parts: a behavioural equation (also called a structural equation in the SEM literature terminology) that links the latent variables to the covariates of interest, and a measurement equation that indicates how the latent variable is reflected in observable indicators. The latent variable can be measured as the principal component using statistical methods like PCA or Confirmatory Factor Analysis. MIMIC models are widely applied in social and clinical psychology, psychometrics, and, to a lesser extent, in economics and political science.

6. Conclusions

This paper has described some of the challenges and important considerations related to measurement and evaluation. We argue that the temptation to focus narrowly on the measurement of the behavioural outcomes of an intervention should be resisted in favour of a detailed appraisal of the drivers underlying these outcomes. This endeavour is key to producing policy evaluations that are useful and that are responsive to the challenges of scaling up successful interventions and the consequential general equilibrium effects.

Fortunately, measurement methodology has recently received renewed attention, both in terms of the methods used to collect data, and in terms of what techniques are used to analyse available data. Furthermore, among economists and in other disciplines, new tools are being developed to measure constructs that, while obviously relevant for understanding policy impacts, have not been measured systematically very often. An obvious example of one such construct is social norms.

Better (and sometimes new) measurements are key to policy evaluations because many of the drivers of behaviour may not be directly or immediately observable. We argue that measurement and theory development go hands in hand, and that they work best when they evolve jointly: new measurement informs more flexible and realistic theories of behaviour and theory informs the construction and design of new measures.

We have discussed three important aspects of measurement development: what to measure, how to measure, and how to create new measures, with illustrative examples from recent measurement innovations. The list of examples we have provided is far from exhaustive: the main purpose of the examples given is to convey more effectively some basic concepts and ideas. It is clear, however, that new developments are happening in the three dimensions we mention above: researchers and policymakers are aware of the possibilities of measuring variables that are relevant for policy analysis. Furthermore, the use of advanced techniques to synthetise effectively available measures has become common. Finally, innovations as regards building new measures, using a variety of techniques and sources (from the use of large administrative data, to the construction of 'lab in the field' experiments, to the use of biomarkers collected in the field) have become more common.

The need to bring about improvements in measurement is present across many disciplines and the process of doing so requires multi- and cross-disciplinary input. Looking at different contexts, it is clear that substantial progress has been made. It is also clear that different disciplines can provide different insights, which makes collaboration and coordination both important and desirable. In this vein, large public initiatives like CEDIL can play an important role in facilitating collaborations and interactions among researchers from different

disciplines, in the standardisation of measures, and in the provision of measurement tools that can be wrested from the exclusive ownership of large private providers of such tools.

References

Advani, A. and Malde, B. (2018) 'Credibly identifying social effects: Accounting for network formation and measurement error', *Journal of Economic Surveys* 32(4), pp. 1016–1044.

Allende, C., Gallego F. and C. Neilson (2019) 'Approximating the Equilibrium Effects of Informed School Choice'. Working paper downloaded at:

https://christopherneilson.github.io/work/equilibrium-informed-school-choice.html

Attanasio, O. P. (2009) 'Expectations and perceptions in developing countries: their measurement and their use', *American Economic Review* 99(2), pp. 87–92.

Attanasio, O. P, F. Cunha and P. Jervis. (2019) 'Subjective Parental Beliefs. Their Measurement and Role', *NBER Working Paper 26516*.

Attanasio, O.P., S. Cattan, E. Fitzsimons, C. Meghir and M. Rubio-Codina (2020) 'Estimating the Production Function for Human Capital: Results from a Randomized Control Trial in Colombia', *American Economic Review* 110(1), pp. 48–85.

Attanasio, O. P., and Kaufmann, K. M. (2017) 'Education choices and returns on the labor and marriage markets: Evidence from data on subjective expectations', *Journal of Economic Behavior & Organization* 140, pp. 35–55.

Barberà, S., Hammond, P. and Seidl, C. (eds.). (2004) *Handbook of Utility Theory: Volume 2 Extensions*. Springer Science & Business Media.

Calvi, R. and Keskar, A. (2021) 'Dowries, resource allocation, and poverty', *Journal of Economic Behavior & Organization* 192, pp. 268-303.

Caplin, A. (2021) 'Economic Data Engineering', *NBER working paper No 29378*.

Cavatorta, E. and Shukri, I. (2022) 'A novel method for measuring stigma in health: evidence from adolescents during COVID-19', *SSRN Working Paper*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4034483

Cavatorta, E. and Groom, B. (2020) 'Does deterrence change preferences? Evidence from a natural experiment', *European Economic Review* 127, 103456.

Cavatorta, E., Zizzo, D., and Daoud, Y. (2020) 'Does exposure to violence affect reciprocity? Experimental evidence from the West Bank', *University of Queensland Working Paper Series* (No. 614).

Chang-Lopez, S., Gertler, P. J., Grantham-McGregor, S., Heckman, J. J., Pinto, R. R. A., Vermeersch, C. M., … and Zanolini, A. (2014) 'Labor market returns to early childhood stimulation: a 20-year followup to an experimental intervention in Jamaica', *Science* 344(6187), pp. 998–1001.

Chen, X., Hong, H. and Nekipelov, D. (2011) 'Nonlinear models of measurement errors', *Journal of Economic Literature* 49(4), pp. 901–37.

Cvitanić, J., Prelec, D., Riley, B. and Tereick, B. (2019) 'Honesty via choice-matching', *American Economic Review: Insights* 1(2), pp. 179–92.

Erdem, T., Keane, M.P. and Öncü, T.S. (2005) 'Learning About Computers: An Analysis of Information Search and Technology Choice', *Quantitative Market and Economics* 3(3), pp. 207–247.

Delavande, A., Giné, X., and McKenzie, D. (2011) 'Eliciting probabilistic expectations with visual aids in developing countries: how sensitive are answers to variations in elicitation design?', *Journal of Applied Econometrics* 26(3), pp. 479-497.

Dominitz, J. (2001) 'Estimation of income expectations models using expectations and realization data', *Journal of Econometrics* 102(2), pp. 165–195.

Dominitz, J. and Manski, C. F. (2004) 'How should we measure consumer confidence?' *Journal of Economic Perspectives* 18(2), pp. 51–66.

Epstein, L. G. and Schneider, M. (2007) 'Learning under ambiguity', *The Review of Economic Studies* 74(4), pp. 1275–1303.

Eriksson, S. and Rooth, D. O. (2014) 'Do employers use unemployment as a sorting criterion when hiring? Evidence from a field experiment', *American Economic Review* 104(3), pp. 1014–39.

Field, E., R. Pande, N. Rigol, S. Schaner and C. Troyer Moore (2021) 'On Her Own Account: How Strengthening Women's Financial Control Impacts Labor Supply', *American Economic Review* 111(7), pp. 2342–75.

Fischbacher, U. and Föllmi-Heusi, F. (2013) 'Lies in disguise—an experimental study on cheating', *Journal of the European Economic Association* 11(3), pp. 525–547.

Galizzi, M. M. and Navarro-Martinez, D. (2019) 'On the external validity of social preference games: a systematic lab-field study', *Management Science* 65(3), pp. 976–1002.

Giustinelli, P. and Pavoni, N. (2017) 'The evolution of awareness and belief ambiguity in the process of high school track choice', *Review of Economic Dynamics* 25, pp. 93–120.

Hopkins, D. and G. King. (2010) 'Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability', *Public Opinion Quarterly*, pp. 1–22.

Karni, E. and Vierø, M. L. (2017) 'Awareness of unawareness: A theory of decision making in the face of ignorance', *Journal of Economic Theory* 168, pp. 301–328.

King, G., Murray, C. J.L., Salomon, J. A. and A. Tandon (2004) 'Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research', *American Political Science Review* 98, pp. 191–207.

King, G, and J. Wand (2007) 'Comparing Incomparable Survey Responses: New Tools for Anchoring Vignettes', *Political Analysis* 15, pp. 46–66.

Hainmueller, J., Hangartner, D. and Yamamoto, T. (2015) 'Validating vignette and conjoint survey experiments against real-world behavior', *Proceedings of the National Academy of Sciences* 112(8), pp. 2395–2400.

Harrison, G. W. and List, J. A. (2004) 'Field experiments', *Journal of Economic Literature* 42(4), pp. 1009–1055.

Levitt, S. D. and List, J. A. (2007) 'What do laboratory experiments measuring social preferences reveal about the real world?', *Journal of Economic Perspectives* 21(2), pp. 153–174.

List, J. A. (2006) 'The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions', *Journal of Political Economy* 114(1), pp. 1–37.

Manski, C. F. (2004) 'Measuring expectations', *Econometrica* 72(5), pp. 1329–1376.

Miguel, E. and Kremer, M. (2004) 'Works: Identifying Impacts on Education and Health in the Presence of Treatment Externalities', *Econometrica* 72(1), pp. 159–217.

Prelec, D. (2004) 'A Bayesian truth serum for subjective data', *Science* 306(5695), pp. 462–466.

Wiswall, M. and Zafar, B. (2021) 'Human capital investments and expectations about career and family', *Journal of Political Economy* 129(5), pp. 1361–1424.

Zizzo, D. J. (2010) 'Experimenter demand effects in economic experiments', *Experimental Economics* 13(1), pp. 75–98.

Centre of Excellence for Development Impact and Learning

www.cedilprogramme.org

@CEDILProgramme

Centre of Excellence for Development Impact